Table of Contents

Recommendations of the Florida Student Growth Implementation Committee: Background and Summary

1. Introduction

Florida is transforming its teacher evaluation system. Under Florida's successful Race to the Top (RTTT) application, districts are committed to participating in the process of developing and using systems of educator evaluation that include student achievement growth measures. The 2011 Florida legislature also passed a law, very closely aligned with Florida's successful RTTT application, requiring that teachers in Florida be evaluated using student achievement data.

The Florida Department of Education (FLDOE) contracted with the American Institutes for Research (AIR) to assist in the development, evaluation, and implementation of a value-added model (VAM) to be used for teacher evaluation. The goal of the project is to provide a fair, accurate, and transparent VAM of teacher effectiveness that districts can incorporate into their teacher evaluation systems to bring about significant educational improvement and to provide useful information about student learning to individual teachers, principals, and other stakeholders.

AIR is working in partnership with the FLDOE and the Student Growth Implementation Committee (SGIC), using a collaborative and iterative process over the next four years to design, develop, analyze, and implement VAMs of student academic performance in Florida public schools at grade levels K–12.

The SGIC made a recommendation to the Commissioner of Education on the value-added model for teachers who teach students in grades and subjects assessed by the Florida Comprehensive Assessment Test (FCAT). As required by the June 1, 2011, deadline established by SB 736, the Student Success Act, Commissioner Eric J. Smith approved a model by announcing his conditional approval of the SGIC's recommendations; however, as part of his conditional approval, Commissioner Smith requested further clarification on the SGIC's "school component" recommendation. After the SGIC clarified that portion of the recommendation, Commissioner Smith fully approved the model on June 8, 2011.

1.1 Summary of Recommendation

The SGIC recommended, and the Commissioner accepted, a value-added model from the class of *covariate adjustment models* (described below). This model begins by establishing expected growth for each student. The expectation is estimated from historical data each year, and it represents the typical growth observed among students who earned similar test scores the past two years and who share several other characteristics. The expected growth increases for students enrolled in more than one course within a specific subject (e.g., mathematics).

The teacher's *value-added score* reflects the average amount of learning growth of the teacher's students above or below the expected learning growth of similar students in the state, using the variables accounted for in the model. In the model recommended by the SGIC, the teacher's *value-added score* is expressed as a sum of two components: one that reflects how much the school's students on average gained above or below similar students in the state (a "school component") and another that reflects how much the teacher's students on average gained above or below similar students within the school (a "teacher component"). The SGIC considered the proportion of the common school component that should be attributed to the teacher and determined that 50 percent of the common school component should be included in the teacher value-added score (a more comprehensive discussion of these issues is provided in Section 5.3). Hence, the recommended final value-added score for teachers is given by

$$Teacher\ Value\ Added\ Score = Unique\ Teacher\ Component + .50 * Common\ School\ Component$$

Ten covariates (variables) are used to establish the expected growth for students:

- The number of subject-relevant courses in which the student is enrolled
- Two prior years of achievement scores
- Students with Disabilities (SWD) status
- English language learner (ELL) status
- Gifted status
- Attendance
- Mobility (number of transitions)
- Difference from modal age in grade (as an indicator of retention)
- Class size
- Homogeneity of entering test scores in the class

The inclusion of these control covariates established expected student scores based on typical growth among students who are similar on these characteristics.

More technically, we can describe the model with the following equation:

$$y_i = \mu + \sum_{g=1}^{M} \delta_g x_g + \sum_{j=1}^{K} \beta_j x_j + \theta_{(k)i} + \omega_{(mk)i} + \varepsilon_i;$$

where y_i denotes the test score for student *i*, δ_g is the coefficient associated with g^{th} prior test score, β_j is the coefficient associated with covariate *j*, θ is the common school component of school *k* assumed $\theta \sim N(0, \sigma_\theta^2)$, ω is the effect of teacher *m* in school *k* assumed $\omega \sim N(0, \sigma_\omega^2)$, and ε is the random error term assumed $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The school and teacher effects were treated as random effects, and the teacher- and school-specific values are empirical Bayes estimates.

The model estimated recognizes that all test scores—both the dependent variable and the independent variables—are measured with finite precision, and the magnitude of precision varies across the scale of test scores. A subsequent technical paper will more fully describe the model and estimation of its parameters.

1.2 Organization of This Report

The remainder of this report proceeds in five sections:

Section 2 summarizes the structure, organization, and role of the SGIC.
Section 3 describes the classes of models initially considered by the SGIC and summarizes the SGIC's decisions.
Section 4 describes the variants of the covariate adjustment models and the difference model considered by the SGIC for further evaluation, along with the specific covariates considered for inclusion.
Section 5 describes the information reviewed by the SGIC to evaluate the models and their ultimate selections.
Section 6 is the appendix.

A technical report to be released in August 2011 will contain all the technical details needed to replicate the selected model.

2. Structure, Organization, and Role of the SGIC

The SGIC is one of eight committees established by the FLDOE to assist with the implementation of RTTT. Over 200 individuals applied to serve on the SGIC. In December 2010, the Commissioner of Education appointed 27 individuals to serve a four-year term on the SGIC.

The members of the SGIC are teachers, principals, parents, union representatives, superintendents, school board members, district administrators, and postsecondary faculty who contribute expertise in various teaching subjects and grades, educational administration at all levels, and measurement and assessment. The SGIC members represent Florida's diversity in culture, community, and region, and they will serve at the appointment of the Commissioner for the four-year term of the project. Sam Foerster, the associate superintendent in Putnam County, serves as the chair of the SGIC. A full membership list is provided in Section 6.1.

2.1. Role of the SGIC

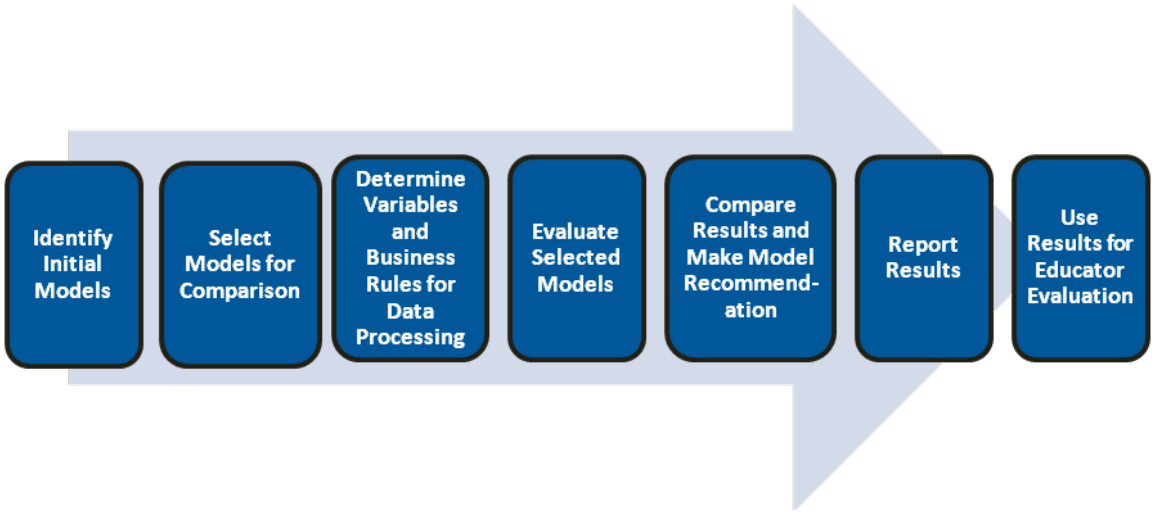
The purpose of the SGIC is to provide input, seek feedback, and present recommendations to the state for the development and implementation of teacher-level student growth models. The SGIC is not responsible for final decisions regarding the adoption of a state model or the district models. The process for providing input, feedback, and recommendations to the state will continue over the four years of the project.

The initial work of the SGIC focused around making a recommendation to the Commissioner of Education on the value-added model to be used for the evaluation of teachers teaching reading and math courses that are assessed with the Florida Comprehensive Assessment Test (FCAT).

Figure 1 illustrates the steps in the process the SGIC followed for selecting a value-added model to recommend to the Commissioner.

To begin the process of selecting a value-added model, illustrated in Figure 1, AIR initially identified eight different value-added models representing the models currently in use in education research and practice. Descriptions, as well as data and policy implications, were presented to the SGIC for each of the models. During the presentation, the SGIC asked questions and began the discussion on the merits of each model for potential use in Florida. At the conclusion of the presentation, the SGIC chair facilitated a discussion that led to a unanimous SGIC decision to have AIR evaluate the differences model and the covariate model with several variants, as described in detail in Section 4. Section 3 provides a detailed description of the eight models initially considered by the SGIC and summarizes the SGIC's decisions on the models selected to move forward in the evaluation process.

Figure 1. Process of Selecting a Value-Added Model



The SGIC also determined which variables to include and data processing rules. In 2011, the Florida legislature passed SB 736, the Student Success Act, which expressly prohibited the use of gender, race/ethnicity, and socioeconomic status as variables in the model. In the same legislation, it was suggested that other variables, such as Students with Disabilities (SWD) status, English language learner (ELL) status, and attendance, be considered as factors in the model. The SGIC discussed the proposed variables and generated a list of additional variables to be considered. The SGIC then discussed each variable individually, determined whether the variable was appropriate for inclusion from a data and policy perspective, and provided a definition for each of the variables. Section 4 describes the variables that were included in the recommended model, as well as those that were considered but were not included in the recommended model. Also included is a summary of the discussion and rationale for the decision.

The SGIC also reviewed the business rules used for processing the data and confirmed that the rules were appropriate. Business rules consist of decisions about student attribution to teachers, how duplicate or missing data is managed, how growth expectations for students taking multiple courses or having multiple teachers are determined, etc. These rules are delineated in the technical specifications paper to be published in August 2011.

Though the law required the selection of a model by the Commissioner on June 1, 2011, the recommendation and selection of a statewide FCAT value-added model does not constitute the end point of the process. Over the next four years, FLDOE and AIR will continue to analyze the value-added model and seek feedback to make adjustments, possibly even before the first year of calculation using the spring 2012 statewide assessment results.

3. Overview of the Classes of Models Considered

This section describes the eight initial value-added models presented to the FLDOE and SGIC for their consideration. AIR did not advocate for or against any particular model. Rather, AIR showcased a variety of models that would allow the SGIC to consider a broad range of model characteristics in selecting the model for Florida. The eight models presented here and to the SGIC were developed to highlight key differences among various approaches to value-added modeling and to allow the SGIC to consider a range of perspectives that exist within the literature and in practice.

Below, we describe the models initially considered by the SGIC and summarize the SGIC's judgments.

3.1. Models Initially Considered by the SGIC

The initial eight models were chosen to represent the diversity found in teacher value-added practice. These models vary across four dimensions:

- The form of the statistical model used to derive the value-added estimates
- The extent to which the models include statistical controls for contextual factors often viewed as outside the control of teachers
- The extent to which past teacher effects remain constant or diminish over time
- The unit of measurement used to represent student achievement (e.g., scale scores versus student percentile ranks)

A brief discussion of each of these dimensions provides context for the differences and similarities among the eight models detailed below.

3.1.1 Form of the Statistical Model

Value-added models run from simple and transparent to quite complex and nuanced. While all VAMs attempt to estimate the systematic component of growth associated with a school or teacher, the complexity of the analysis used to accomplish this task varies considerably. In general, to measure growth, models control for the prior achievement of students in some way. The complexity of the model is determined by how models account for prior achievement, how

the model estimates value-added scores of school and teacher effects, and assumptions about the sustainability of school and teacher effects. While there are many different statistical approaches to value-added modeling, AIR grouped the approaches into two main classes for presentation to the SGIC: (1) typical learning path models and (2) covariate adjustment models.

Typical Learning Path Models

AIR dubbed the first class of models *typical learning path* models (more technically known as general longitudinal mixed-effects models). These models assume that each student has a "typical learning path." Absent effects of schools or teachers, each student's expected performance is a given number of points above the conditional average, with that number being estimated from multiple years of data. This number can be thought of as a student's *propensity to achieve*. The model posits that schools and teachers can alter this learning path, increasing or decreasing the student's path relative to the state mean.

One characteristic of these models is that they do not directly control for prior achievement. In fact, the control can be more accurately described as controlling for typical student achievement. As additional data accumulate, a student's propensity to achieve can be estimated with more accuracy. This characteristic implies that, with each passing year, better estimates become available for past years (because the student's typical learning path is estimated with increased precision over time).

Learning path models must make some assumptions about *how* teachers or schools impact a student's propensity to achieve. Different analysts make different assumptions about the durability of a teacher's effect on a student's typical learning path. In Sanders' Tennessee Value-Added Assessment System (TVAAS) model, teacher effects are assumed to have a permanent impact on students. McCaffrey and Lockwood (2008) estimated a model that relaxes this assumption and lets the data dictate the extent to which teacher effects decay over time. Indeed, in an experiment in Los Angeles, Kane et al. (2008) found that teacher effects appeared to dissipate over the course of about two years.

Covariate Adjustment Models

The second class of models, covariate adjustment models, directly controls for prior student scores. These models can treat teacher effects as either fixed or random. Unlike the first class of models, covariate adjustment models directly introduce prior test scores as predictors in the model. Thus, covariate models directly control for past achievement whereas typical learning path models control for a "propensity to achieve" over time, which is estimated from past and current achievement. To obtain unbiased results, covariate adjustment models must account for measurement error introduced by the inclusion of model predictors (prior student achievement). Two widely used methods for accounting for the measurement error in regression analyses include modeling the error directly (as in structural equation models or errors-in-variables regression) and an instrumental variable approach, which uses one or more variables that are assumed to influence the current year score, but not prior year scores, to statistically purge the measurement error from the prior year scores.

3.1.2 Statistical Controls for Contextual Factors

Both learning path and covariate adjustment models can vary in the extent to which they control for contextual factors (e.g., student, classroom, and school characteristics). The previous section described how each of the main classes of models controls for prior student achievement. Controlling for prior student achievement is both qualitatively different from controlling for other student characteristics and statistically necessary to obtain valid estimates of teacher value-added because students are not sorted randomly into districts, schools, and classes. Rather, there are purposive selection mechanisms that cause certain teachers to encounter certain students in their classrooms. These mechanisms include parent selection of schools and teachers; teacher selection of schools, subjects, and sections; and principal discretion in assigning certain students to certain teachers. All of these selection factors cause significant biases when not addressed in models that estimate teacher value-added.

Unbiased estimates of teacher value-added require that factors that influence both selection of students into particular classes *and* current year test scores be statistically controlled. Many value-added models assume that the only selection factor that is relevant to the outcome (the student's posttest score) is the student's prior test score. Such models assert that effectively controlling for that score leaves the student assignment to classrooms conditionally independent of the posttest score. This, of course, assumes the use of appropriate statistical methods that facilitate unbiased control for prior test scores. Others models incorporate controls for additional variables thought to influence selection and outcomes.

The empirical evidence is mixed on the extent to which student characteristics other than score histories remain correlated with test scores after controlling for prior test scores. Ballou, Sanders, and Wright (2004) find that controlling for student-level characteristics makes little if any significant difference in model estimates, and McCaffrey et al. (2004) report similar findings under most conditions. These findings are consistent with the view that durable student characteristics associated with race, income, and other characteristics are already reflected in prior test scores, such that controlling for the prior test scores controls for any relevant impact of the factors proxied by the measured characteristics. In contrast, when student factors are aggregated to school or classroom levels, they sometimes reveal a significant residual effect (Raudenbush, 2004; Ballou, Sanders, and Wright, 2004). In other words, school or classroom characteristics (e.g., high percentage of students with IEPs) may explain additional variance in students' posttest scores independently beyond students' individual characteristics accounted for by their prior test scores. Raudenbush interprets this as potentially reflecting a peer effect—an interpretation with which Sanders takes issue. The significance of the effect of aggregate student characteristics at the classroom or school level on student learning should not be dismissed. If schools with highly disadvantaged student populations are systematically served by less effective teachers, the data would reveal significant associations between aggregated characteristic measures and growth or value-added. To the extent that any of these characteristics are related to true effectiveness, true teacher effectiveness really does vary with student characteristics, and correlated variation of estimated teacher value-added is not the consequence of uncontrolled selection bias but rather a reflection of these true differences in teacher effectiveness.

3.1.3 Durability of Teacher Effects

As noted above, typical learning path models require an assumption about the durability of the impact of teachers on a student's learning path. Popular value-added models vary in their assumptions about how and whether teacher effects decay over time. In Sanders' Tennessee Value-Added Assessment System (TVAAS) model, teacher effects are assumed to have a permanent impact on students. McCaffrey and Lockwood (2008) estimated a model that relaxes this assumption and lets the data dictate the extent to which teacher effects decay over time. In an experiment in Los Angeles, Kane et al. (2008) found that teacher effects appeared to dissipate over the course of about two years.

Covariate models generally do not require assumptions about the durability of teacher effects. This is because they explicitly establish expectations based on prior achievement by including prior test scores as a covariate, rather than the more abstract "propensity to achieve" estimated in learning path models.

3.1.4 Unit of Measurement for Student Achievement

A growing number of states have adopted variants of the Colorado growth model (Betebenner, 2008). This model is entirely normative, replacing test scores (which are used with the models described previously) with in-state percentile ranks. This model has the advantage of not relying on a potentially flawed vertical scale. However, this model sacrifices the ability to establish anything but normative criteria. For example, if students across the state made little progress in middle school mathematics, a student who remained at the 50th percentile from one year to the next could conceivably be losing proficiency. In particular, that student, while keeping up with similar peers near the middle of the score distribution, could be forgetting what he or she had learned previously. Thus, the Colorado growth model examines students' growth relative to their peers rather than absolute growth in their own learning.

Betebenner's (2008) model uses quantile regression to estimate student growth curves in the percentile metric. While there are many approaches to estimating the quantile functions, Betebenner and colleagues use a parametric method based on an assumed curve.

3.1.5 Initial Models Presented to the SGIC

Eight initial models were presented to the SGIC for the committee's consideration. These models varied on the four modeling specifications described above—namely, the form of the statistical model, the use of contextual control variables, the durability of teacher effects, and the unit of measurement. Below we describe the major characteristics of each model presented to the SGIC in April 2011. The first four models are examples of typical learning path models, and the second four models are examples of covariate adjustment models.

3.1.6 Learning Path Models

Model 1: Similar to the Sanders Model

Model 1 is similar to Sanders' Tennessee Value-Added Assessment System (TVAAS) model, which assumes that teachers have a permanent impact on students. This model is often referred to as a layered model because the impact of teachers is thought to "layer" on top of the

impact of prior teachers, permanently altering a student's propensity to learn. Table 1 describes the specifications for model 1, the Sanders model relative to the four considerations described above.

Table 1. Model 1 Specifications

Dimension	Model Specifications
Form of the Statistical Model	Typical learning path model that estimates the amount of learning growth systematically associated with the teacher, controlling for a student's typical performance over time. Teacher effects are modeled as random effects.
Use of Contextual Control Variables	No contextual control variables are used.
Durability of Teacher Effects	Teacher effects are assumed not to decay with time.
Unit of Measurement	Test scores (i.e., interval measures of student achievement)

Model 2: Similar to the Rand Model

The second model presented to the SGIC was similar to the one developed by McCaffrey and Lockwood (2008) and is typically referred to as a variable persistence model (referring to the possibility that teacher effects do or do not persist over time). The model is similar to the Sanders model except that it does not assume that teacher effects on a student layer upon each other year after year. Instead, the impact of each teacher is thought to dissipate over time. Table 2 summarizes the specifications of this model relative to the four dimensions.

Table 2. Model 2 Specifications

Dimension	Model Specifications
Form of the Statistical Model	Typical learning path model that estimates the amount of learning growth systematically associated with the teacher, controlling for a student's typical performance over time. Teacher effects are modeled as random effects.
Use of Contextual Control Variables	No contextual control variables are used.
Durability of Teacher Effects	Teacher effects on a student can vary over time and are directly estimated from the data (rather than being assumed to be constant, or persist, over time).
Unit of Measurement	Test scores (i.e., interval measures of student achievement)

Model 3: Hybrid Model 1

Model 3, the first of two hybrid models presented to the SGIC, is nearly identical to the Rand model, with the notable exception that teacher effects are estimated as fixed rather than random effects. Fixed effects explicitly model teacher effects, estimating a parameter for each teacher,

but they can be computationally burdensome. In contrast, random-effects models simply estimate the mean and variance of the (presumably normal) distribution of teacher effects. It is not until the model estimation is complete that the estimated distribution of effects and the observed data for teachers are combined to infer a specific effect for each teacher. As with the Rand model, teacher effects are allowed to decay over time. The specifications of this model are presented in Table 3.

Table 3. Model 3 Specifications

Dimension	Model Specifics
Form of the Statistical Model	Typical learning path model that estimates the amount of learning growth systematically associated with the teacher, controlling for a student's typical performance over time. Teacher effects are modeled as fixed effects.
Use of Contextual Control Variables	No contextual control variables are used.
Durability of Teacher Effects	Teacher effects on a student can vary over time and are directly estimated from the data (rather than being assumed to be constant, or persist, over time).
Unit of Measurement	Test scores (i.e., interval measures of student achievement)

Model 4: Hybrid Model 2

The second hybrid model is also similar to model 2 (the Rand model), with the exception that the model controls for additional characteristics of the student, school, or class. Model 4 models random teacher effects whose durability over time is estimated directly from the data. Table 4 summarizes the model specifications for model 4.

Table 4. Model 4 Specifications

Dimension	Model Specifics
Form of the Statistical Model	Typical learning path model that estimates the amount of learning growth systematically associated with the teacher, controlling for a student's typical performance over time. Teacher effects are modeled as random effects.
Use of Contextual Control Variables	Student characteristics and contextual variables are included in the model.
Durability of Teacher Effects	Teacher effects on a student can vary over time and are directly estimated from the data (rather than being assumed to be constant, or persist, over time).
Unit of Measurement	Test scores (i.e., interval measures of student achievement)

3.1.7 Covariate Adjustment Models

Model 5: Similar to the Meyer Model

The fifth model, similar to the Meyer model (Meyer, 1992; Meyer, 2010), estimates student growth as the amount a teacher's typical student learns above and beyond that which is typical for students with similar characteristics, controlling for prior achievement. This model is the first of four covariate adjustment models—that is, these models include previous achievement (i.e., test scores) as model predictors. This type of model is generally considered more transparent and easier to implement than the general longitudinal random effect models described in models 1 to 4. AIR proposed estimating this model with an errors-in-variables approach (as opposed to the instrumental variables approach typically used with the Meyer model). AIR suggested this method because it supports, but does not require, the inclusion of student characteristic variables (which are required in Meyer's approach for calculating estimates of prior scores). This particular version of the Meyer model includes student characteristic variables (model 6 is a similar model but excludes student characteristic variables). Table 5 summarizes the model specifications for model 5.

Table 5. Model 5 Specifications

Dimension	Model Specifics
Form of the Statistical Model	Covariate adjustment model that directly controls for prior student achievement by including these variables as terms in the regression model.
Use of Contextual Control Variables	Student characteristics and contextual variables are included in the model.
Durability of Teacher Effects	Not applicable
Unit of Measurement	Test scores (i.e., interval measures of student achievement)

Model 6: Hybrid Model 3

Model 6 estimates the same model as model 5 (the Meyer model)—that is, it estimates the amount of learning growth systematically associated with a teacher but excludes student characteristic variables from the model. A comparison of models 5 and 6 would show whether including student-level characteristics changed the estimates of teacher effects. Table 6 summarizes the model specifications for model 6.

Table 6. Model 6 Specifications

Dimension	Model Specifics
Form of the Statistical Model	Covariate adjustment model that directly controls for prior student achievement by including these variables as terms in the regression model.
Use of Contextual Control Variables	No contextual control variables are used.
Durability of Teacher Effects	Not applicable
Unit of Measurement	Test scores (i.e., interval measures of student

	achievement)

Model 7: The Differences Model

The differences model is a variation of the model that AIR implemented statewide in Florida with the Foundation for Excellence in Education to select teachers for the Excellence in Teaching Award. Because this model does not estimate the relationship between prior and current year test scores, the model is not biased by the errors-in-variables problem inherent in covariate adjustment models. In contrast, this model uses difference scores that can be conceived as the measure obtained by subtracting the prior test score from the current one. This is mathematically equivalent to fixing the regression coefficient associated with the prior test score to one. Model 7 is the most transparent model proposed. It allows for estimates of the systematic learning growth associated with a teacher, while requiring little in the way of sophisticated statistics or complex estimation. If this type of model were to yield estimates similar to those of more complex approaches, it may be the preferable model although a more complex model may be preferred to explicitly account for factors that are important to policymakers. The specifications for model 7 are presented in Table 7.

Table 7. Model 7 Specifications

Dimension	Model Specifics
Form of the Statistical Model	This is a covariate adjustment model. However, in contrast to models 5 and 6, this model fixes the model coefficient associated with prior achievement at 1.
Use of Contextual Control Variables	No contextual control variables are used.
Durability of Teacher Effects	Not applicable
Unit of Measurement	Test scores (i.e., interval measures of student achievement)

3.1.8 Quantile Regression Model

Model 8: Similar to the Colorado Model

Model 8 is different from all the previous models in that its objective is to develop normative tables for student growth (similar to the height and weight charts for children one may see in a pediatrician's office). Furthermore, this model uses student percentile rank within the student's grade as the dependent variable; all previous models have used the interval measure of a student's scaled test score as the outcome measure. For this model, teacher effectiveness is measured as the typical growth percentile of a teacher's students (relative to what would have been expected given expected growth curves). Effective teachers are identified as those whose students grow more than what would have been expected.

Normative measures of student growth estimate growth in students' achievement relative to their peers. While it is possible to relate these normative estimates to actual scale score, doing

so eliminates most of the motivation for estimating the inherently normative model. Table 8 summarizes the model specification for model 8.

Table 8. Model 8 Specifications

Dimension	Model Specifics
Form of the Statistical Model	This model is fit as a quantile regression model that predicts student growth conditional on a student's past location within the distribution of student scores (percentile rank).
Use of Contextual Control Variables	No contextual control variables are used.
Durability of Teacher Effects	Not applicable
Unit of Measurement	In-state percentile ranks

The eight models were selected to demonstrate the variability in approaches to value-added modeling and (if recommended by the SGIC) to provide empirical evidence regarding the extent to which various model specifications (e.g., the inclusion or exclusion of student characteristics) would impact teacher effects. Table 9 provides a summary of the specifications across the eight models.

Table 9. Summary of Specifications of Initial Proposed Models

Model	Form of Statistical Model	Teacher Effects	Contextual Control Variables	Durability of Teacher Effects	Unit of Measurement
1	Typical learning path	Random	None	Sustained	Interval Scale
2	Typical learning path	Random	None	Variable Persistence	Interval Scale
3	Typical learning path	Fixed	None	Variable Persistence	Interval Scale
4	Typical learning path	Random	Included	Variable Persistence	Interval Scale
5	Covariate adjustment	Random	Included	N/A	Interval Scale
6	Covariate adjustment	Random	None	N/A	Interval Scale
7	Covariate adjustment	Fixed	None	N/A	Interval Scale
8	Covariate adjustment	Random	None	N/A	Percentile Rank

3.2 SGIC Considerations

The SGIC proceeded by ruling out models that did not meet their requirements. Discussions, codified in a series of resolutions, first ruled out the typical learning path models and then the

percentile rank models, leaving the committee with the decision to consider a variety of covariate adjustment models, including the difference model.

3.2.1. Eliminating Typical Learning Path Models

The SGIC began by considering the different classes of models—typical learning path models versus the covariate adjustment class of models. Discussion at the meeting favored the covariate adjustment models over the typical learning path models for the following reasons:

- SGIC members were less comfortable with the more abstract nature of the control for student achievement in the typical learning path model.
- SGIC members were not comfortable with the notion that better estimates of past teacher effect become available in the future, potentially altering past measures of teacher effect.
- The SGIC recognized the greater reliance that these models have on the interval properties of the measurement scale

First, recall that the typical learning path models control for a student's "propensity to achieve" rather than directly controlling for his or her past performance. Past performance reveals the amount above or below a student's peers the student is expected to perform in the absence of an exceptional teacher or school intervention. Value-added modeling estimates the extent to which teachers alter this learning path. However, some models assume that the teacher's impact is permanent and does not diminish over time, while others assume that teacher effects decay over time. These different assumptions have important implications for teachers' value-added scores. For example, a year with a particularly good teacher might forever increase expectations for a student, even for a student who might already be high-performing relative to his or her peers. Given these assumptions and the less tangible foundation for student expectations, the SGIC found learning path models less desirable.

Second, typical learning path models gain information about each student's learning path over time. Because the path is assumed constant over time (except for the impact of teachers or schools), more accurate estimates for past years become available each year. This implies that teacher evaluations could require ongoing revisions of prior assessments as these estimates become increasingly precise with each additional year of available data. These revisions could be problematic. For example, the SGIC spent some time discussing the hypothetical possibility that data collected subsequently would suggest that the previous dismissal of a teacher (based on prior value-added estimates) would be unwarranted after those estimates were updated with new data from subsequent years. These considerations again raised significant concerns about the use of learning path models.

Third, the SGIC recognized that the FCAT measurement scale, like all measurement scales, is neither perfect nor perfectly equated across grades. The learning path models rely quite heavily on the assumptions about equal measurement intervals across the range of the scale and across grades. Because of these concerns and those highlighted above, the SGIC elected not to use learning path models.

3.2.2. Eliminating Percentile Rank Models

The SGIC considered the percentile rank model but did not favor its inherently normative nature. The percentile rank growth models generally rank students' growth and use aggregates of these percentile ranks as a measure of teacher or school effectiveness.

The SGIC was concerned that the standards expressed in percentile ranks would preclude the possibility that all teachers ever meet the standard. The growth standards would not have a straightforward expression on scale scores and would not easily connect to established proficiency levels.

For these reasons, the SGIC voted to eliminate percentile rank models from consideration.

3.2.3. Retaining Covariate Adjustment Models

The SGIC voted to retain covariate adjustment models, including the differences model that can be considered a special case of such models. These variants and the considerations that led to them are described in the next section.

4. Specific Model Variants Estimated and Considered by the SGIC

The SGIC wanted to evaluate the impact of four potential modeling decisions:

> What is the impact of including two prior years of achievement rather than one prior year of achievement?
> What is the impact of estimating the common school component (*school effect*) within the teacher value-added model?
> What is the impact of including different subsets of covariates?
> How do the more complex covariate adjustment models compare with the difference model?

One of the SGIC's critical endeavors involved identifying the covariates to include in the model (question 3). We discuss this below, followed by a description of the model variants estimated, evaluated, and presented to the SGIC.

4.1 Selection of Covariates to Include in the Model

VAMs are designed to mitigate the influence of differences among students in teachers' entering classes. Covariates intend to "level the playing field" so that schools and teachers do not have advantages or disadvantages simply as a result of the students who attend a school or are assigned to a class.

The most important control, theoretically and empirically, is prior student achievement scores. Students are not randomly sorted into schools or classroom. There are significant differences across schools and classrooms in the entering proficiency of students. A variety of mechanisms contribute to this phenomenon, including parent selection of schools and teachers; teacher selection of schools, subjects, and sections; and principal discretion in assigning certain students to certain teachers.

Unbiased estimates of teacher value-added do not require random assignment of students into classrooms. Instead, the effects of selection are mitigated when factors included in the model are those that (a) are not accounted for by pretest scores and (b) are associated with posttest scores after controlling for pretest scores.

The 2011 Florida legislature explicitly prohibited using the variables gender, race/ethnicity, and socioeconomic status in the model and suggested that variables such as Students with Disabilities status, English language learner status, and attendance be considered.

At the April 4 and 5 SGIC meeting at the University of Central Florida, the SGIC generated a list of potential variables for discussion. At the meeting, it was determined that specific student characteristics, including Students with Disabilities (SWD) status, gifted status, English language learner (ELL) status, and attendance, would be evaluated in these models as determined and defined by the SGIC. Several additional variables were discussed on an April 14 SGIC webinar that resulted in the inclusion of class size, age, mobility, school effect, and homogeneity of class.

When considering variables for inclusion in the model, the SGIC used the following framework to guide the discussion:

Data are available and accurate.
Discussion on variable inclusion:
- Is it in the teacher's control?
- Is it measured already by another variable?
- Is it explained by pretest data?

Possible definitions

Below is a list of variables considered and approved by the SGIC that were *included* in the value-added models.

Students with Disabilities (SWD) status
English language learner (ELL) status
Gifted status
Attendance
Mobility
Difference from modal age in grade (as an indicator of retention)
Class size
Homogeneity of entering test scores in the class
School effect

Below is a list of variables considered by the SGIC that were *excluded* from the value-added models:

Response to Intervention (RTI) level (a method of early academic intervention to assist struggling students)
Foster care status
Rural schools

Homework
Teacher attendance
Teacher experience
Migrant status
Homeless status
School grades
Availability of resources
Course complexity
Discipline and referral action

4.1.1 Discussion and Summary of Variables

Although the SGIC considered a wide range of possible variables for inclusion in the VAMs, many variables were excluded for a variety of reasons. Specifically, four variables including Response to Intervention (RTI) level, foster care status, rural schools, and homework were not available when AIR conducted the evaluation for the results provided to the SGIC in May. Some SGIC members suggested that the FLDOE should consider beginning a consistent collection of these data, and, if that occurs, these variables should be reconsidered. In some cases, the data were available (e.g., teacher experience) although there were concerns about accuracy and/or whether the variable could be appropriately defined for use in the value-added models.

The SGIC elected to exclude teacher attendance and teacher experience from the models, but the committee recommended that AIR evaluate the relationship between teacher attendance and teacher value-added scores as part of the results presented to the SGIC in May 2011 for consideration. Additionally, after considering other variables, the SGIC eliminated homework, homeless status, and migrant status. Specifically, the SGIC expressed the belief that homework is within the teacher's influence. Homeless status was believed to be too closely related to the legislatively prohibited socioeconomic status. Further, rather than measuring migrant status, the SGIC thought that including mobility would provide sufficient control.

The SGIC briefly considered school grades or other school resources (e.g., computer resources) as a covariate in the VAM but elected to exclude these measures because the data available were insufficient. The SGIC also considered including a measure of course complexity but decided not to include it in the model. Additionally, although data were available for discipline and referral action, the SGIC excluded this variable because of inconsistency in reporting and because some SGIC members believed that the variable is teacher-controlled.

4.1.2 Testing SGIC-Approved VAMs

The SGIC granted AIR the authority to test all the discussed models and to develop variations of the models (with different subsets of control variables) to address the sensitivity of teachers' value-added scores to different model specifications.

AIR configured the variables into three sets including a set that excluded all covariates other than baseline performance ("none"), a set that included an abbreviated list of covariates ("few"), and a set that included the full list of approved covariates ("many"). Table 10 summarizes the variables in each of these groups. The variables in the abbreviated covariate list were those that

the SGIC were most committed to including. The full covariate list included these same variables, as well as the rest of the variables of interest to the SGIC. Together, these three groups provide a test of the sensitivity of teachers' value-added scores when different subsets of covariates are included in the model.

Table 10. Covariates Included in Each Set

No Covariates Other than Baseline Performance Data ("None")	Abbreviated Covariate List ("Few")	Full Covariate List ("Many")
The number of subject-relevant courses in which the student is enrolled	The number of subject-relevant courses in which the student is enrolled	The number of subject-relevant courses in which the student is enrolled
Prior years of achievement scores	Prior years of achievement scores	Prior years of achievement scores
	Students with Disabilities (SWD) status	Students with Disabilities (SWD) status
	English language learner (ELL) status	English language learner (ELL) status
	Gifted status	Gifted status
	Attendance	Attendance
		Mobility (number of school transitions)
		Class size
		Homogeneity of entering test scores in the class
		Difference from modal age in grade (as an indicator of retention)

4.2. Overview of Models Compared

To address the four questions outlined at the beginning of this section, AIR estimated six variants of the covariate adjustment model and the differences model. The particular variants were designed to offer key comparisons to address the questions raised by the SGIC. These variants are described in Table 11.

Table 11. Summary of Covariate Adjustment Models Estimated

Number of Scores from Prior Years	Other Covariates Included	Teacher Only	Teacher and School
1 Prior Year Only	None		Model 1
	Few	Model 2	
2 Prior Years	None		Model 3
	Few	Model 4	Model 5
	Many		Model 6

This configuration, along with the differences model, allowed for independent comparisons that addressed each of the questions. Table 12 summarizes the model comparisons that supported inquiry into each of the guiding issues.

Table 12. Summary of Model Comparisons Addressing Each Question

Question	Models Compared
What is the impact of including two prior years of achievement rather than one prior year of achievement?	Models 1 and 3
What is the impact of including separately estimating the common school component (*school effect*) within the teacher value-added model?	Models 2 and 4
What is the impact of including different subsets of covariates?	Models 4, 5, and 6
How do the more complex covariate adjustment models compare with the differences model?	Any model compared to the differences model

5. Information Reviewed by the SGIC to Evaluate Models

In this section we describe the criteria by which the SGIC evaluated the various models in order to narrow their choice. These criteria were established for the SGIC, and empirical data regarding each model were presented so the SGIC could judge each model in terms of the specific criteria.

The SGIC was provided with two lenses by which they could compare the models:

1. Empirical data showing how the models compared on multiple criteria
2. Impact data showing some real-world relationships of the VAM results with other factors

In this section we describe the results that were provided to the SGIC and summarize their reactions.

5.1. Characteristics of the FCAT Assessment

Upon initial review of the FCAT data, we observed two key phenomena that supported the choice of a covariate adjustment model by grade. First, when examining the vertical scale, we observed abnormally large changes in the mean performance of students within a grade over time (e.g., see grade 8 reading scores in Exhibit 1). This suggests that linking error may be large and could be conflated with teacher effects. A possible consequence is that any model that explicitly uses the student-level gains as the outcome variable could exaggerate (or understate) true gains in learning as a result of the linking error associated with the test scale.

Second, the equal interval property of the scale across grades was questionable. For instance, students in grade 10 reading have, on average, much smaller learning gains than students in grade 7 reading (see Exhibit 2). If the model estimated teacher effects simultaneously for all grades, grade 7 teachers could appear to produce higher value-added than grade 10 teachers as a result of the test scale and not as a result of real instructional practices.

Both of these issues led AIR and the SGIC to focus attention on the covariate adjustment models, which do not require a vertical scale and may be more robust to the year-to-year changes in test performance and for the VAM to be computed on a grade-by-grade basis.

Exhibit 1. Reading FCAT Average Scale Score Year by Year

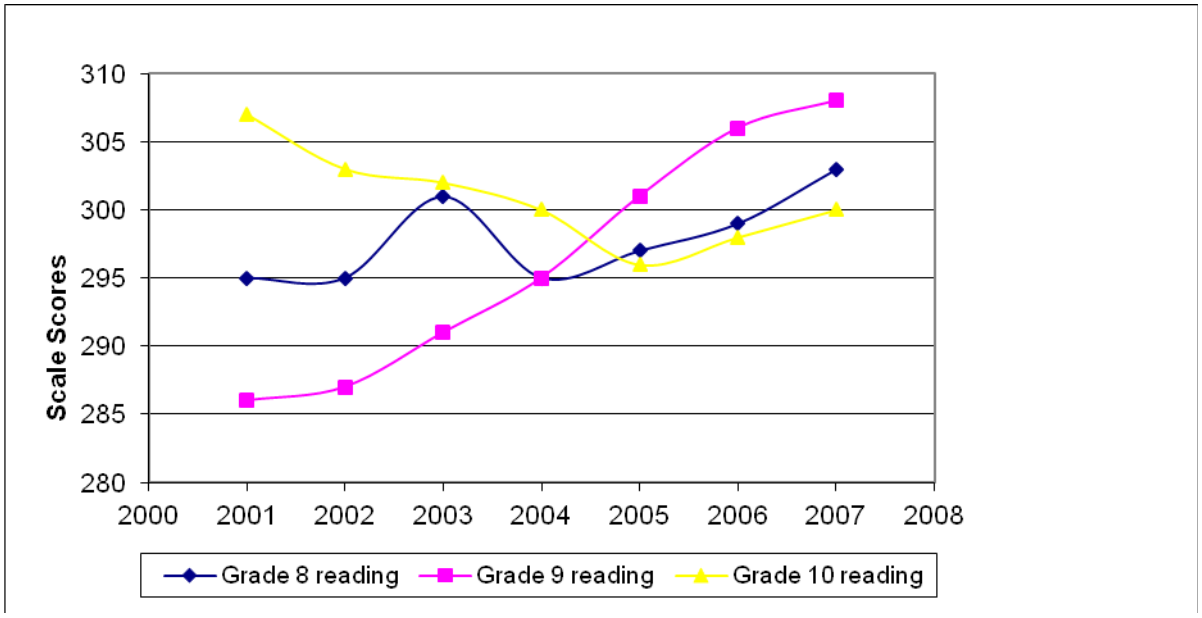
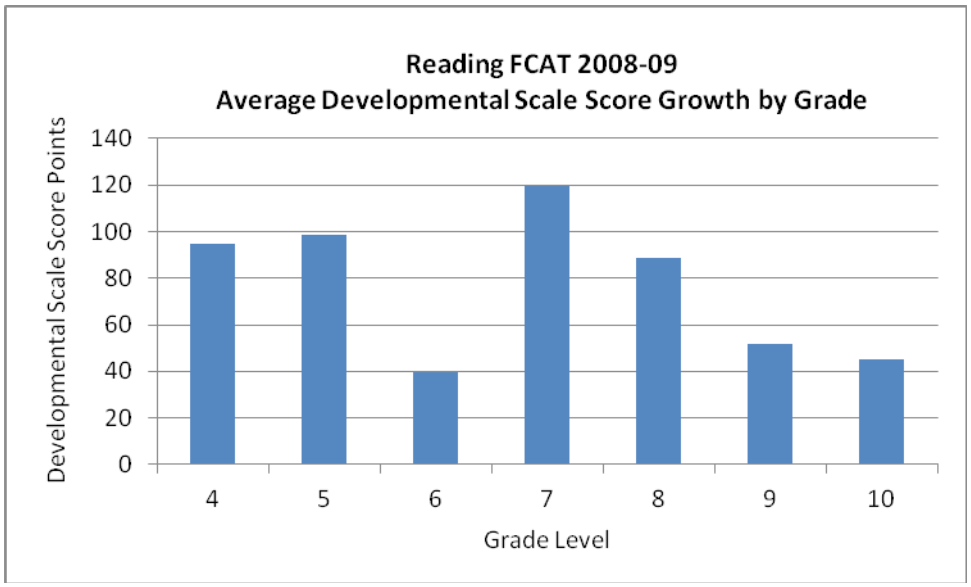


Exhibit 2. Reading FCAT 2008-09 Average Developmental Scale Score Growth by Grade



5.1.1. Summary of Simulations

Prior to implementing any of the models, AIR completed a series of simulations to examine the bias and adequacy of the model-based standard errors. Although these simulations will be extensively detailed in a technical report, here we note that for each model we created 200 data sets with known model parameters. We tested each of the models to assess whether they recovered the true values of the parameters and whether they produced standard errors that captured the real-world variability of the parameters. For all models, we observed no bias, and

we found that the model-based standard errors were adequate representations of the observed sampling variability.

5.1.2. Similar Composition of Classrooms

The FCAT data provide the links of students to classrooms that are necessary to support the implementation of the VAM. However, because some teachers have classrooms that consist of virtually all of the same students as another class, the estimates of fixed teacher effects demonstrated some instability. When the composition of two classrooms is identical, the matrix used for VAM computation becomes singular (meaning that the VAM cannot compute teacher effects). As the proportion of students shared between two teachers' increases, the estimates become less stable. When a random effects framework is used these issues do not affect estimation (because individual teacher effects are not estimated, but instead are calculated post hoc after estimation).

5.1.3. Precision of the Teacher Effects

AIR also examined the statistical precision with which the various models estimated teacher effects. Obviously, teacher effects measured with greater precision are preferred over teacher effects measured with less precision, other things being equal. To assess the precision of the estimates, we examined the standard errors of the teacher effects. Exhibits 3 and 4 show the distribution of teacher effect standard errors in grade 7 math and reading across all models estimated. The black dot in the plot is the median standard error, and the box plot shows the 5^{th}, 25^{th}, 75^{th}, and 95^{th} percentiles. These visual displays are useful for making judgments on which model yields smaller standard errors, thus indicating teacher effects measured with greater precision.

The plots show at least two important characteristics of the models the SGIC considered. First, note that the differences model is presented on a different scale. The differences model estimates larger absolute-value effects and has correspondingly larger standard errors.

Second, the models that include two prior test scores always provide more precise estimates of the teacher effects. This pattern is observed in both reading and math. Based on this criterion, the SGIC was able to narrow consideration of the models to those including two prior years of scores.

Exhibit 3. Distribution of the Standard Error of Estimated Teacher Effects Under Eight Models in Math FCAT Grade 7

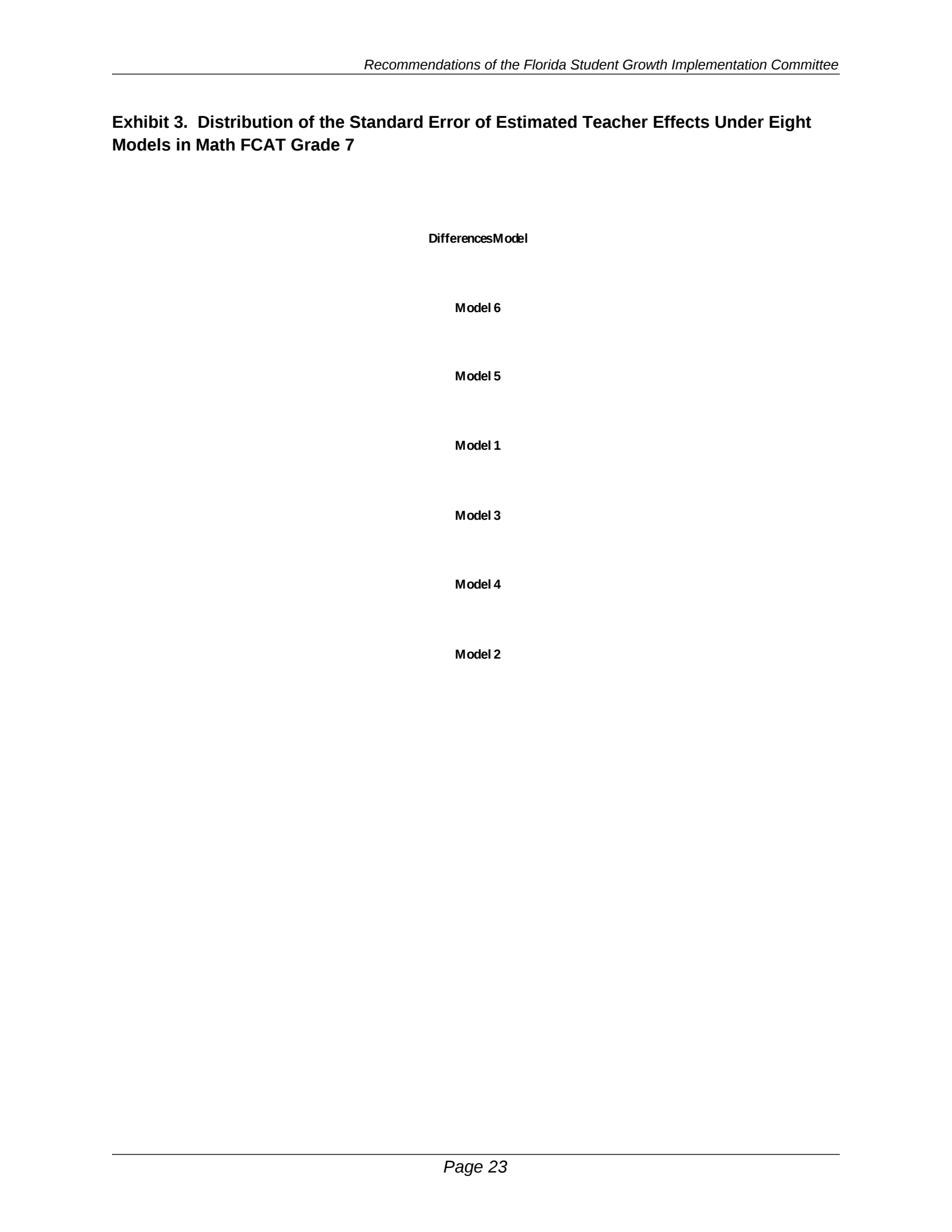
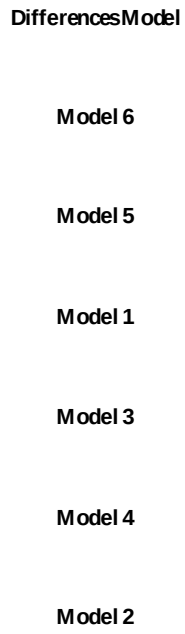


Exhibit 4. Distribution of the Standard Error of Estimated Teacher Effects Under Eight Models in Reading FCAT Grade 7



5.2. Impact Information

The preceding discussion outlines the criteria by which the SGIC evaluated each VAM. Additionally, AIR provided the SGIC with impact data to illustrate the potential observed consequences of the various models, showing the extent to which expectations differed between student groups, as well as the correlations between the value-added scores and the various teacher factors. Specifically, the impact data included a review of these factors:

Whether expectations for learning differ between
- English language learners (ELL) and non-ELL students;
- gifted and non-gifted students; and
- students with different prior test scores.

Relationship of teacher effects with
- teacher experience;
- teacher attendance (number of absences);
- percentage of Students with Disabilities in a classroom;
- percentage of ELL students taught in a classroom; and
- highest teacher degree obtained.

Results presented to the SGIC demonstrate the following:

1. The models all showed a larger growth expectation for ELL students than for non-ELL students. AIR's and the SGIC's working hypothesis was that ELL students have lower initial scores because of their limited ability to comprehend the math and reading text. After one year of immersion, they seemingly perform better on the test because of their improved language comprehension. That is, they show larger gains presumably as a result of their improved ability to understand the test questions.

2. Non-gifted students showed larger growth expectations than gifted students. The SGIC believed that this could result from the test scale where gains at the top end may be more difficult to achieve (i.e., a ceiling effect).

3. The data showed a negative correlation between growth expectations and prior test score. As a consequence, students in the lowest quartile tend to have slightly larger growth expectations than students in higher quartiles.

4. All of the models produced estimates that had little or no correlation with other factors examined, including teacher absenteeism and experience. Two nontrivial correlations were observed: Teachers teaching high proportions of ELL students and teachers with high proportions of very low-performing students were more likely to have high value-added scores (correlations are on the order of –.1 to –.2).

Although these results provided added context for interpreting the VAM results presented to the SGIC, none of these findings raised significant concerns. The SGIC concluded that these findings were plausible and reflected real-world phenomena.

5.3. Attribution of the Common School Component of Growth to Teachers

The VAM applied to the FCAT data decomposes total growth above expectation among a teacher's students into a common school component and a teacher component. While all parameters are estimated simultaneously, it is a useful heuristic to consider the levels separately. First, student-level prior test scores (i.e., the lags) and the covariates are used to establish a statewide expectation. This expectation is the score a student is expected to have, given his or her prior test score history and characteristics.

However, schools exhibit differential amounts of growth. The model cannot differentiate whether these differences are due to independent factors at the school (e.g., particularly effective leadership) or simply due to the sorting of high-growth teachers into some schools rather than others. We refer to this as the "common school component" of student growth. The common school component therefore describes the amount of learning that is typical for students in each school that differs from the statewide expectation.

Teacher effects can be interpreted as deviations from the common school component of learning. In the models where the common school component is not estimated, the common school component is implicitly attributed to the teacher.

Whether or not to estimate the common school component and teacher effects was a source of significant discussion for the SGIC, and it is a source of significant discussion in the value-

added literature. If school level factors exert an independent influence on student learning, then ignoring the school component implicitly attributes the effect of those forces to teachers. Such factors might include particularly effective (or ineffective) school leadership, curriculum, resources, peer influences, or community factors. Committee members considered the possibility that such factors, if attributed to the teachers, might create an incentive for teachers to avoid struggling schools.

The committee also considered the possibility that all systematic between-school differences might be due entirely to the different mix of teachers at the school. If that were the case, *not* attributing the school component to the teachers could create adverse incentives. For example, working together to improve the teaching of all teachers would result in a higher school component, for which the teachers would be denied credit.

The committee agreed that the in the real world the school component probably reflected a mix of teacher impact and school-level factors that are independent of the teachers.

After significant discussion, as well as with a second follow-up meeting, the SGIC determined that some of the school effect should be attributed back to teachers. The proportion allocated back was put to vote and agreed upon by the SGIC as 50 percent. Hence, teacher value-added scores are determined using the following calculation:

*Teacher Value Added Score = Unique Teacher Component + .50 * Common School Component*

This formula simply recognizes that some of the school effect is a result of teacher actions within their schools and that they should receive some credit in their overall value-added effects.

5.4. Conclusion

In total, the SGIC considered an extensive amount of data regarding the VAMs applied to the FCAT data and its impact on estimating teacher effects. By using the previously described criteria and impact data as well as policy discussions and professional judgment, the SGIC decided to use Model 6, which included teacher and school effects, two prior test scores, and the full set of additional covariates. The SGIC's final recommendation is based on a comprehensive review of the data and their professional insights into which model will best serve the policy aims set forth in state law.

6. Appendix

6.1. Florida's Student Growth Implementation Committee (SGIC) Members

***Sam Foerster*, Chair,** Associate Superintendent, Putnam
***Sandi Acosta*,** Teacher (6th and 7th Science), Dade
***Ronda Bourn*,** Consortium Administrator
***Anna Brown*,** Representative for Superintendent MaryEllen Elia, Hillsborough
***Joseph Camputaro*,** Teacher (Elementary/Reading), Lee
***Julia Carson*,** Teacher (HS AP History, Geography), Volusia
***Cathy Cavanaugh*,** Postsecondary, UF
***Doretha Wynn Edgecomb*,** School Board, Hillsborough
***Gisela Field*,** District Administrator – Assessment, Dade
***Stacey Frakes*,** Teacher (3rd – 5th ESE), Madison
***Arlene Ginn*,** Teacher (7th Science), Orange
***Stephanie Hall*,** School-based Administrator (ES), Brevard
***Lavetta B. Henderson*,** Postsecondary, FAMU
***Eric O. Hernandez*,** Teacher (Honors Math), Dade
***Linda J. Kearschner*,** Parent, Pinellas
***Latha Krishnaiyer*,** State PTA
***John le Tellier*,** Teacher (Music), Marion
***Nicole Marsala*,** Teacher (8th History), Broward
***Lisa Maxwell*,** Local Union, Broward
***Lawrence Morehouse*,** Business
***Jeff Murphy*,** District Administrator - Student Services, Virtual School
***Maria Cristina Noya*,** School-based Administrator (HS), St. Lucie
***Pam Stewart*,** Assistant Superintendent, St. Johns
***Lance J. Tomei*,** Postsecondary, UCF
***Gina Tovine*,** District Administrator – HR, Levy
***Lori Westphal*,** Teacher (ESE), Lake
***Tamar E. Woodhouse-Young*,** Teacher (High School Math), Duval

7. References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66.

Betebenner, D. W. (2008). *A primer on student growth percentiles* (Technical report). Dover, NH: National Center for the Improvement of Educational Assessment.

Kane, T.J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* Unpublished. Cambridge, MA: Harvard University.

Kane, T., & Staiger, D. (2008, April). *Are teacher-level value-added estimates biased: An experimental validation of non-experimental estimates*. Paper delivered at the National Conference on Value-Added Modeling, Madison, WI.

McCaffrey, D., Sass, T., & Lockwood, J. R. (2008, April). *The intertemporal stability of teacher effect estimates*. Paper delivered at the National Conference on Value-Added Modeling, Madison, WI.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Meyer, R. H. (1992). Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency. Discussion Paper No. 96-92, University of Wisconsin-Madison, Institute for Research on Poverty.

Meyer, R.H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197–223). Washington, DC: National Academy Press.

Meyer, R. H. (2010). *Value-added models and the next generation of assessments.* Princeton, NJ: Educational Testing Service.

Raudenbush, S. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.